Effects of Worldwide Population Subdivision on *ALDH2* Linkage Disequilibrium

Raymond J. Peterson,[1,2,3] David Goldman,[1] and Jeffrey C. Long[1]

[1]Laboratory of Neurogenetics, National Institute on Alcohol Abuse and Alcoholism, National Institutes of Health, Bethesda, Maryland 20892-8110 USA; [2]Department of Anthropology, Pennsylvania State University, University Park, Pennsylvania 16802 USA

The effect of human population subdivision on linkage disequilibrium has previously been studied for unlinked genes. However, no study has focused on closely linked polymorphisms or formally partitioned linkage disequilibrium within and among worldwide populations. With an emphasis on population subdivision, the goal of this paper is to investigate the causes of linkage disequilibrium in *ALDH2*, the gene that encodes aldehyde dehydrogenase 2. Haplotypes for 756 people from 17 populations across five continents were estimated by maximum-likelihood from genotypes at six closely linked *ALDH2* nucleotide substitutions. Linkage disequilibrium was partitioned into three components: within populations, among populations within continents, and among continents. It was found that population subdivision among continents had a larger and more disparate effect on linkage disequilibrium than subdivision among local populations. Further, linkage disequilibrium did not increase with population divergence as predicted by a simple model. Rather, the patterns of linkage disequilibrium were complicated because of the interplay of a near absence of recombination, the linkage disequilibrium that existed prior to the divergence of modern humans, subsequent mutation, population subdivision, random genetic drift, and perhaps natural selection. These results suggest that simple models may not well predict patterns of linkage disequilibrium in human populations.

Linkage disequilibrium is the nonrandom association of alleles at different loci. In an ideal population at equilibrium, linkage disequilibrium is predicted to approach zero at a rate dependent on the recombination fraction. However, linkage disequilibrium can be generated by genetic drift (Hill and Robertson 1968), population subdivision (Nei and Li 1973), natural selection (Lewontin 1964), and mutation (Ohta 1982a,b). Because of this, it is not surprising that complicated patterns of linkage disequilibrium are observed in human populations (Jorde et al. 1994; Lewontin 1995; Clark et al. 1998).

Despite the complexity of observed patterns, several expectations have emerged. One is that linkage disequilibrium is expected to peak near a disease gene when the disease allele is rare (Ajioka et al. 1997). Application of this principle led to the positional cloning of the genes for cystic fibrosis (Kerem et al. 1989) and diastrophic dysplasia (Hästbacka et al. 1992, 1994). Another expectation is that linkage disequilibrium between a frequent disease allele and alleles at marker loci may be best preserved in a small, constant-sized population (Laan and Pääbo 1997, 1998; Terwilliger et al. 1998). However, recent theoretical work challenges this view for frequent alleles (Lonjou et al. 1999). With respect to population subdivision, while linkage disequilibrium is expected to vary among subdivisions of finite size, the average among subdivisions is expected to be zero (Hill and Robertson 1968). Finally, the variance of linkage disequilibrium is expected to increase with population subdivision and to decrease with migration (Ohta 1982a).

For pairs of unlinked genes the effect of population subdivision on linkage disequilibrium has been studied in the Tecumseh, Michigan population (Sinnock and Sing 1972) and within South American Indian villages (Smouse and Neel 1977; Smouse et al. 1983). In both studies an excess number of statistically significant values were attributed to the populations being recently founded by migrants from source populations that differed in allele frequency. In addition, Smouse and colleagues (Smouse and Neel 1977; Smouse et al. 1983) found that the effect of population subdivision on linkage disequilibrium was greater among clusters of villages than among local villages.

For pairs of closely linked polymorphisms, three studies have examined linkage disequilibrium in worldwide samples. Castiglione et al. (1995) investigated alleles at a dinucleotide short tandem repeat polymorphism (STRP) and two restriction site polymorphisms (RSPs) in *DRD2*. Tishkoff et al. (1996, 1998) examined alleles at a pentanucleotide STRP and an Alu deletion polymorphism in *CD4*, and a trinucleotide STRP, an Alu deletion polymorphism, and two RSPs in *DM*. All three studies found that African populations had many haplotypes and low levels of linkage disequilibrium. In contrast, nonAfrican populations had a

[3]**Corresponding author.**
E-MAIL peterson@ncifcrf.gov; FAX (301)846-1909.

9:844–852 ©1999 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/99 $5.00; www.genome.org

subset of the African haplotypes and almost complete linkage disequilibrium. These results were attributed to a founder event at the time modern humans emigrated from Africa.

The goal of this paper is to investigate the causes of worldwide linkage disequilibrium in *ALDH2*, the gene that encodes aldehyde dehydrogenase 2. *ALDH2* is located on chromosome 12q24.2 (Raghunatan et al. 1988) and spans 44 kb (Hsu et al. 1988). Haplotypes were estimated from alleles at six biallelic sites within *ALDH2* (Fig. 1) that were genotyped in 756 people from 17 populations across five continents (Peterson et al. 1999). *ALDH2* has a dominant deficiency allele that is frequent in, but private to, Asia (Yoshida et al. 1984). The deficiency allele is of interest because natural selection in the form of conferring resistance to parasite infection may have preserved this allele in Asia (Ikuta et al. 1986; Goldman and Enoch 1990; R.J. Peterson, D. Goldman, and J.C. Long, in prep.).

Table 1. Frequency of the Variant Allele (×1000) at Six Sites in 17 Worldwide Populations

		Site					
Populations	2N	1	2	3	5	6	12
Biaka	102	98	225	225	0	235	0
Africa	102	98	225	225	0	235	0
Cambodian	48	146	188	188	146	188	146
Chinese	94	106	170	170	309	170	298
Japanese	98	163	102	102	276	102	286
S. Korean	80	162	175	175	113	175	113
Taiwanese	86	116	233	233	267	233	267
Black Thai	100	140	230	230	60	23	60
Asia	506	138	182	182	200	182	200
CEPH	64	766	219	219	0	219	0
Finn	82	793	195	195	0	195	0
Swede	90	844	144	144	0	144	0
Europe	236	805	182	182	0	182	0
Cheyenne	102	647	88	88	0	88	0
Mayan	100	510	110	110	0	11	0
Navajo	92	815	76	76	0	76	0
Pima	90	556	233	233	0	233	0
N. America	384	630	125	125	0	125	0
Karitiana	98	571	194	194	0	194	0
R. Surui	88	943	11	11	0	11	0
Ticuna	98	541	235	235	0	235	0
S. America	284	676	151	151	0	151	0
World	1512	465	165	165	67	165	67

RESULTS

Allele and Haplotype Frequency

The allele frequencies at each site and in each population are shown in Table 1. Examination of the multisite homozygotes and single-site heterozygotes yielded seven directly observed haplotype states. The maximum-likelihood frequency estimates of these haplotypes and their jackknife standard errors are tabulated in Table 2. Below, haplotype states are given within brackets. A 1 represents the reference allele, and 2 represents the variant allele. The alleles are ordered by site where the sites are in the order 1, 2, 3, 5, 6, and 12. For brevity, each haplotype state is designated by a number and the letter H. The site and haplotype numbers are from Peterson et al. (1999).

Three haplotypes had worldwide distribution: H1 [111111], H2 [211111], and H3 [122121] (Fig. 2). Of note, the frequencies of H1 and H2 were nearly reversed in the African Biaka and in Europeans and the variant alleles at sites 2, 3, and 6 usually co-occurred. Although H4 [111212] was private to Asia, it attained a frequency of 25% in the Chinese, Taiwanese of Chinese descent, and Japanese. H4 carried the deficiency allele as well as the usually co-occurring variant at site 5. The high frequency of H4 in Asia appears to have come about largely at the expense of H1. The combined frequency of H1 and H4 in Asia is 67.9%, almost identical to the 66.7% frequency of H1 in the African Biaka. The remaining haplotypes were observed in single copy only: H6 [111121] in the African Biaka, H8 [111211] in the Chinese and H9 [111112] in the Japanese.

Population Divergence

The variants at the six sites naturally formed three groups of sites: site 1; sites 2, 3, and 6; and sites 5 and 12. Sites within each group yielded nearly identical allele frequency distributions and fixation indices. Fixation indices (Wright 1978), or *F*-statistics, measure population divergence as the among group proportion of the total allele frequency variance. To avoid redundancy the *F*-statistics are not reported individually but rather for each group of sites. F-statistics were calculated for local populations relative to continental average, continental average to world-

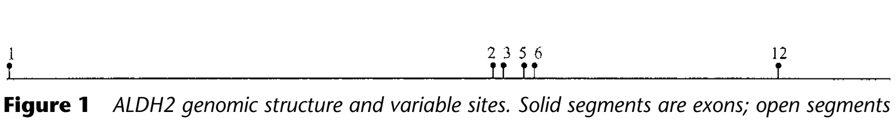

Figure 1 *ALDH2 genomic structure and variable sites. Solid segments are exons; open segments are introns. Numbers indicated the variable sites that were genotyped in the worldwide survey. (●) Nucleotide substitutions.*

Table 2. Estimated Frequency of Unique *ALDH2* Haplotypes (×1000) and Jackknife SE in 17 Worldwide Populations

		Haplotypes[a]						
Populations	2N	H1	H2	H3	H4	H6	H8	H9
Biaka	102	667	98	225	0	10	0	0
		(52)	(32)	(40)	(0)	(10)	(0)	(0)
AFRICA	102	667	98	225	0	10	0	0
		(48)	(29)	(41)	(0)	(10)	(0)	(0)
Cambodian	48	542	137	176	145	0	0	0
		(114)	(92)	(81)	(48)	(0)	(0)	(0)
Chinese	94	415	106	170	298	0	11	0
		(49)	(37)	(39)	(47)	(0)	(11)	(0)
Japanese	98	449	163	102	276	0	0	0
		(51)	(36)	(28)	(46)	(0)	(0)	(10)
S. Korean	80	550	163	175	112	0	0	0
		(60)	(46)	(40)	(36)	(0)	(0)	(0)
Taiwanese	86	384	116	233	256	0	0	0
		(52)	(35)	(46)	(47)	(0)	(0)	(0)
Black Thai	100	570	140	230	60	0	0	0
		(51)	(35)	(46)	(23)	(0)	(0)	(0)
ASIA	506	481	136	179	198	0	2	2
		(22)	(15)	(17)	(18)	(0)	(2)	(2)
Ceph	64	16	766	218	0	0	0	0
		(16)	(54)	(49)	(0)	(0)	(0)	(0)
Finn	82	12	793	195	0	0	0	0
		(13)	(47)	(47)	(0)	(0)	(0)	(0)
Swede	90	11	844	145	0	0	0	0
		(11)	(34)	(36)	(0)	(0)	(0)	(0)
EUROPE	236	13	805	182	0	0	0	0
		(7)	(26)	(25)	(0)	(0)	(0)	(0)
Cheyenne	102	265	647	88	0	0	0	0
		(49)	(53)	(28)	(0)	(0)	(0)	(0)
Mayan	100	380	510	110	0	0	0	0
		(52)	(55)	(32)	(0)	(0)	(0)	(0)
Navajo	92	109	815	76	0	0	0	0
		(34)	(39)	(27)	(0)	(0)	(0)	(0)
Pima	90	211	556	233	0	0	0	0
		(43)	(53)	(44)	(0)	(0)	(0)	(0)
N. AMERICA	384	245	555	200	0	0	0	0
		(22)	(25)	(17)	(0)	(0)	(0)	(0)
Karitiana	98	235	571	194	0	0	0	0
		(37)	(47)	(40)	(0)	(0)	(0)	(0)
R Surui	88	45	943	11	0	0	0	0
		(23)	(23)	(12)	(0)	(0)	(0)	(0)
Ticuna	98	224	541	235	0	0	0	0
		(34)	(44)	(41)	(0)	(0)	(0)	(0)
S. AMERICA	284	173	676	151	0	0	0	0
		(22)	(28)	(21)	(0)	(0)	(0)	(0)
WORLD	1512	302	446	183	66	1	1	1
		(12)	(13)	(9)	(6)	(1)	(1)	(1)

[a] With 1 as the reference allele, the haplotype configurations are H1: 111111; H2: 211111; H3: 122121; H4: 111212; H6: 111121; H8: 111211; and H9: 111112.

wide average, and local populations to worldwide average. In the following, S indexes local populations, C indexes continental averages, and T indexes the worldwide average.

At site 1, allele frequency differences among local populations resulted in an F_{SC} of 8% (Table 3). Allele frequency differences among continents resulted in an F_{CT} of 37%. The divergence among all sub-populations (F_{ST}) was 42%. Because of the almost one-to-one correspondences between haplotype frequency and variant allele frequency, these *F*-statistics can also be explained in terms of the haplotype frequency variation. At site 1, the *F*-statistics largely reflect the H1 and H2 frequency reversal in the African Biaka and the Europeans.

Sites 2, 3, and 6 contrasted the frequency of H3 with H1, H2, and H4. Reflecting the low frequency variation of H3 among populations F_{SC} was 3%, F_{CT} was 0%, and F_{ST} was 3%. Sites 5 and 12 contrasted H4 with H1, H2, and H3. Here, F_{SC} was 12%, F_{CT} was 17%, and F_{ST} was 27%. These latter *F*-statistics were due entirely to the restriction of H4 and the deficiency allele to Asia. Treating the haplotypes as multiple alleles at a single locus, the haplotypic F_{SC} was 5.9%, F_{CT} was 24.4%, and F_{ST} was 28.8%. Here too the low frequency variation of H3 among populations contributed little to these values.

As the jackknife standard errors indicate (Table 3), the confidence intervals for F_{SC} often overlap zero but those of F_{CT} or F_{ST} usually do not. This result indicates that subdivision among continents has played the more important role in divergence of allele and haplotype frequencies. As indicated by their standard errors, F_{ST} values for site 1 (42%) and for sites 5 and 12 (27%) were significantly larger than the 10%—15% F_{ST} values usually reported for RSPs (Bowcock et al. 1991; Jorde et al. 1995). Such large values may be due to random genetic drift, natural selection, or both.

Two-Site Linkage Disequilibrium Analysis

The linkage disequilibrium coefficient ($D_{A_1B_1}$) compares haplotype frequency ($P_{A_1B_1}$) with the product of the allele frequencies (p_{A_1} and q_{B_1}). That is, $D_{A_1B_1} = P_{A_1B_1} - p_{A_1}q_{B_1}$ (Weir 1996). Hereafter *D* is given without any

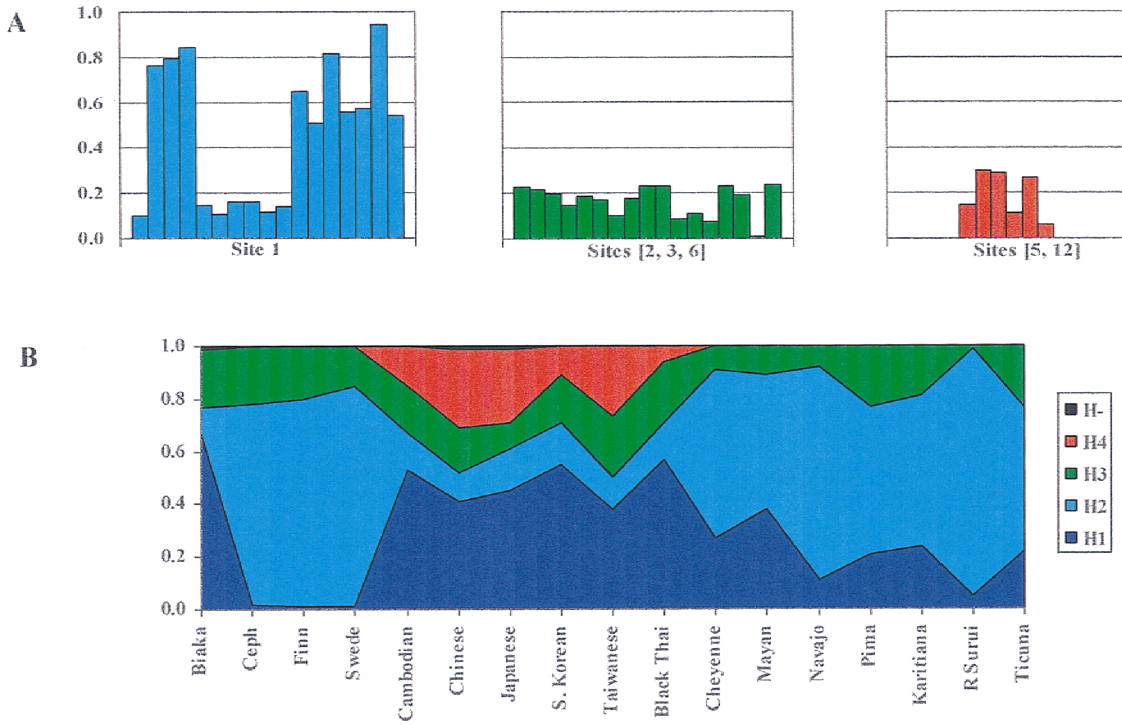

Figure 2 Allele and haplotype frequencies. (*A*) Frequency of the variant (minor) allele at each site in each population. The alleles are colored to match the corresponding haplotypes depicted in *B*. The order of populations is the same as *B*. (*B*) Hpalotype frequencies in each population. H− = H6 + H8 + H9.

subscripts when the argument pertains to a pair of alleles at any two sites. Because *D* depends on allele frequency, it was normalized to the maximum that it could have been given the allele frequencies: D′ = $D/D_{\max}$ (Lewontin 1964). It was also normalized to the following correlation coefficient (Hill and Robertson 1968):

$$r_{A_1B_1} = D_{A_1B_1} \Big/ \sqrt{p_{A_1}p_{A_2}q_{B_1}q_{B_2}}$$

Within each population *D*′ was −1.0 or +1.0. This outcome is consistent with the fact that the variability at all six sites was essentially carried on only four haplotypes. The approximate variance of *D*′ is 0 when *D*′ = −1.0 or +1.0 (Zapata et al. 1997), making these *D*′ values statistically significant. The correlation coefficient (*r*) results are given in Table 4. For pairings that involve site 1, *r* was relatively low except in the Europeans. The Europeans were unique in that only two haplotypes dominated the frequency spectrum. Low *r* values were also observed for the 2, 3, 6 versus 5, 12 pairing in Asia. In contrast, and concordant with the usual co-occurrence of alleles at sites 2, 3, and 6, and at sites 5 and 12, the *r* values for pairings within these groups were at or near unity.

In a subdivided population, the total linkage disequilibrium (D_T) can be partitioned into additive components using a hierarchical model (Nei and Li 1973). Here D_T was partitioned into $D_W + D_{SC} + D_{CT}$ where D_W is the average linkage disequilibrium within populations, D_{SC} is the linkage disequilibrium among local populations, and D_{CT} is the linkage disequilibrium among continents. D_W, D_{SC}, and D_{CT} were then normalized to D_T to obtain d_W, d_{SC} and d_{CT}. Interestingly, D_{SC} and D_{CT} depend solely on allele frequency differences among groups (Nei and Li 1973). Consequently allelic divergence among populations can increase, decrease, or leave unchanged linkage disequilibrium.

Table 3. Fixation Indices and Jackknife Standard Errors

Sites[a]	F_{SC} (%)	F_{CT} (%)	F_{ST} (%)
1	8 (5)	37 (7)	42 (8)
2,3,6	3 (2)	0 (1)	3 (1)
5,12	12 (5)	17 (5)	27 (4)
Haplotypes	6 (3)	24 (3)	29 (4)

[a] See text for details.

The partitioning of worldwide *ALDH2* linkage disequilibrium revealed that linkage disequilibrium within populations (d_W) usually accounted for most of the total linkage disequilibrium (Table 5). In addition, the effect of population subdivision was greater and more disparate among continents (d_{CT}) than among local populations (d_{SC}). Specifically, d_{SC} ranged from just 1% to 6% whereas d_{CT} ranged from −10% to 70%. The d_{CT} values can be explained by the fact that large among group values require large allele frequency differences at both sites of the two-site haplotype (Sinnock and Sing 1972). As indicated by the jackknife standard errors, all of the d_{SC} and d_{CT} estimates were significantly different from 0. It can be concluded that population subdivision, both among local populations and among continents, had a significant effect on the worldwide linkage disequilibrium.

The within-population r^2 values (computed from Table 4) were plotted against the haplotypic F_{SC} and F_{CT} values (Fig. 3). These values were then compared with Hill and Robertson's (1968) model of population divergence, which predicts that linkage disequilibrium increases with F_{ST} (broken line). The *ALDH2* r^2 values ranged from well below to almost as far above the predicted line as was possible. Clearly, the model of Hill and Robertson (1968) did not fit the *ALDH2* data well. This result is perhaps not surprising given that this island model assumed a large population initially in linkage equilibrium, equality of population sizes, and the absence of mutation and natural selection. The evolutionary history of *ALDH2* likely violates several if not all of these assumptions. Furthermore, Hill and Robertson's model assumed that the product of effective population size multiplied by the recombination rate was large. This is not likely for the closely linked sites surveyed here.

Table 5. Partition of Worldwide *ALDH2* Linkage Disequilibrium: Estimates and Jackknife Standard Errors

	Linkage disequilibrium components			
Two-site systems[a]	d_W (%)	d_{SC} (%)	d_{CT} (%)	D_T
1–2,3,6	88	6	6	−0.08
	(2.0)	(1.0)	(3.0)	
2,3,6	96	3	1	0.14
	(0.3)	(0.3)	(0.2)	
1–5,12	29	1	70	−0.03
	(1.0)	(0.3)	(2.0)	
2,3,6–5,12	104	6	−10	−0.01
	(4.0)	(2.0)	(4.0)	
5,12	81	5	14	0.06
	(1.0)	(1.0)	(1.0)	

[a]See text for details.

Table 4. Two-Site *ALDH2* Linkage Disequilibrium Correlation Coefficients (*r*) in 17 Worldwide Populations

	Two-site haplotype systems[a]				
Population	1–2,3,6	2,3,6	1–5,12	2,3,6–5,12	5,12
Biaka	−0.18	0.98	—	—	—
Africa	−0.18	0.98	—	—	—
Cambodian	0.01	1.00	−0.17	−0.20	1.00
Chinese	−0.16	1.00	−0.23	−0.30	0.97
Japanese	−0.15	1.00	−0.27	−0.21	0.97
S. Korean	−0.20	1.00	−0.17	−0.18	1.00
Taiwanese	−0.20	1.00	−0.22	−0.33	1.00
Black Thai	−0.22	1.00	−0.01	−0.14	1.00
Asia	−0.17	1.00	−0.20	−0.24	0.99
Ceph	−0.96	1.00	—	—	—
Finn	−0.96	1.00	—	—	—
Swede	−0.96	1.00	—	—	—
Europe	−0.96	1.00	—	—	—
Cheyenne	−0.42	1.00	—	—	—
Mayan	−0.36	1.00	—	—	—
Navajo	−0.60	1.00	—	—	—
Pima	−0.61	1.00	—	—	—
N. America	−0.49	1.00	—	—	—
Karitiana	−0.57	1.00	—	—	—
R. Surui	−0.43	1.00	—	—	—
Ticuna	−0.60	1.00	—	—	—
S. America	−0.61	1.00	—	—	—

[a]See text for details.

DISCUSSION

A striking pattern of *ALDH2* haplotypic variation was the maximal linkage disequilibrium and corresponding low number of haplotypes. While the number of haplotypes that segregate at a locus depends on historical effective population size and natural selection, combinatorics show that there are 2^s possible haplotype states from *s* biallelic sites. From a related perspective, the cladistic model of haplotype evolution predicts that $s + 1$ haplotypes must have existed in evolutionary history to establish variability at each site. These primary haplotypes create a network of haplotypes that differ from each other by single mutational steps (Long et al. 1990). Some or all of the remaining $2^s - s - 1$ haplotype states could exist in a population because of recombination.

At *ALDH2*, $2^6 = 64$ haplotype states are possible, but only seven states were observed and only four were frequent. At least $6 + 1 = 7$ one-step haplotypes must have existed in evolutionary history. However, it is impossible that the seven observed haplotypes comprise the primary set. H3 differs from H1 at three sites, and the two intermediate one-step haplotypes are completely, or essentially, missing. Whereas H6 provides an intermediate link at one of the steps, only a single copy was observed and it may have arisen by recombination. Similarly, H4 differs from H1 at two sites. Although H8 and H9, each observed in single-copy,

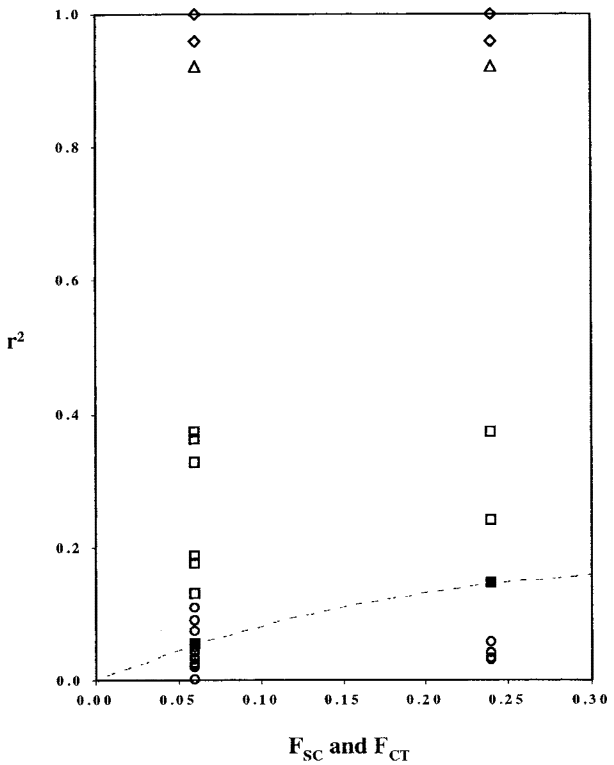

Figure 3 Contrast of population divergence with variance of linkage disequilibrium. F_{sc} (0.06) is contrasted with the within-population r^2 values, and F_{CT} (0.24) is contrasted with the continental r^2 values. (Broken line) Relationship between F_{st} and r^2 predicted from the model of Hill and Robertson (1968). (■) Predictions for F_{st} = 0.06 and F_{st} = 0.24. (○) Site 1 vs. 2, 3, 6 (Africa and Asia), site 1 vs. 5, 12 (Asia); and sites 2, 3, 6 vs. 5, 12 (Asia) (□) Site 1 vs. 2, 3, 6 (Americas); (△) site 1 vs. 2, 3, 6 (Europe); (◇) sites 2, 3, 6 (Africa, Europe, Asia, Americas); site 5 vs. 12 (Asia) (note that each ◇ accounts for several data points; see text for details).

connect H4 and H1 by single steps, at least one of these haplotypes must have been formed by recombination, and it is possible that both were. Thus, three of the seven one-step haplotypes were essentially, or entirely, missing. In humans, one-step haplotypes are frequently missing, as evidenced by *β-globin* (Harding et al. 1997) and *NF1* (Jorde et al. 1993) haplotype phylogenies.

The number of segregating haplotypes at *ALDH2* may be due to natural selection at *ALDH2* (R.J. Peterson, D. Goldman, and J.C. Long, in prep.) or selection on a closely linked gene. A coalescent analysis of the *ALDH2* haplotype phylogeny suggests that, given a neutral model, the age of the deficiency allele is expected to be 149,000 (35,000–416,000) years (R.J. Peterson, D. Goldman, and J.C. Long, in prep.). Such an ancient apparent age rivals the origin of modern humans and predates the colonization of Asia. This suggests that natural selection has increased the frequency of the deficiency allele in Asia faster than expected under a neutral model, and directional selection can reduce the number of haplotypes at a locus. Alternatively, the low number of *ALDH2* haplotypes could be the result of a population bottleneck that is recent relative to the mutation rate.

Because only one African population was sampled, a complete understanding of the African versus non-African patterns of *ALDH2* haplotype variation awaits the sampling of more African populations. Speculatively, the fact that the African Biaka shared the worldwide pattern of linkage disequilibrium at sites 2, 3, and 6 suggests that the pattern arose before the divergence of modern humans and has not subsequently decayed. An absence of a strong out-of-Africa effect at *ALDH2* is hinted by the similarity of the H1 frequency in the Biaka with the H1 + H4 frequency in Asia. Interestingly, while the variants at sites 5 and 12 were in complete linkage disequilibrium, their presence only in Asia indicates a recent Asian origin and perhaps natural selection. Thus, the African versus non-African pattern at *ALDH2* may contrast with *DRD2*, *CD4*, and *DM*. At these latter loci non-Africans had higher linkage disequilibrium and segregated a subset of African haplotypes (Castiglione et al. 1995; Tishkoff et al. 1996; 1998). However, this pattern was not as extreme at *DRD2*. This suggests that an out-of-Africa effect had less effect at *DRD2* than at *CD4* and *DM*.

These distinct patterns of linkage disequilibrium have several explanations. An out-of-Africa founder event may by chance have had less effect at *ALDH2*, or natural selection could have been stronger. In contrast to the other loci, the *ALDH2* haplotypes did not comprise STRP alleles. The STRP mutation rate is several orders of magnitude higher than the nucleotide substitution rate (Weber and Wong 1993). Because of this, STRPs may better resolve recent human evolution. Whatever the explanation, these divergent results suggest that patterns of linkage disequilibrium vary across the human genome.

In addition, the *ALDH2* evidence suggests that the pattern of linkage disequilibrium at any set of closely linked sites may depend on the pattern of linkage disequilibrium that existed in ancestral populations. Because each gene may have had a unique pattern of linkage disequilibrium in ancestral populations, the effect of population subdivision at individual genes may be idiosyncratic. This insight contrasts with the situation of unlinked genes (Smouse and Neel 1977; Smouse et al. 1983). Because unlinked genes have a low covariance of allele and haplotype frequency, a particular pattern of allelic divergence can occur from many different starting haplotype distributions. Thus, while the effect of population divergence on *ALDH2* is likely to be locus dependent, the effect of population divergence on unlinked genes is likely to be independent of the particular set of loci studied.

The importance of ancestral linkage disequilibrium to current patterns has recently received theoretical treatment (Lonjou et al. 1999). These investigators showed that for ancient polymorphisms linkage disequilibrium is largely determined by regional founders, whereas subsequent demography has little effect. This theory was supported by data at the MNSs, RHCE, and *CD4* loci. Of implication to genetic epidemiologists, and contrary to current belief (Terwilliger et al. 1998), is that isolates may actually be less advantageous than large populations for linkage disequilibrium studies (Lonjou et al. 1999).

Another important insight is that patterns of linkage disequilibrium may vary within a set of closely linked sites. At *ALDH2*, the effect of population subdivision varied greatly depending on the groups of sites that were compared. This complicated pattern reiterates that linkage disequilibrium among populations is not a simple function of population divergence (Nei and Li 1973). Distinct patterns of linkage disequilibrium were also observed among tightly linked sites in the *AI–CIII* apolipoprotein gene region (Thompson et al. 1988). Specifically, linkage disequilibrium was found between two flanking RSPs but not with an internal RSP. In this case, the power to detect linkage disequilibrium was low because the major allele of each flanking RSP occurred with the rare allele of the internal RSP (Thompson et al. 1988). These results suggest that patterns of linkage disequilibrium in a gene region may not be fully described by analyzing a single pair of sites. Rather, the proper characterization of linkage disequilibrium may require the examination of alleles at several sites. This same conclusion was reached in a linkage disequilibrium analysis of 88 variable sites in the human lipoprotein lipase gene (Clark et al. 1998).

Population subdivision clearly affected the worldwide pattern of *ALDH2* linkage disequilibrium. Linkage disequilibrium among local populations and among continents was significantly different from zero. The magnitude of this effect was greater among continents than among local populations. The present study augments the original one-level model of population subdivision (Sinnock and Sing 1972; Nei and Li 1973) by extending it to a second level. Because it was found that the effects of population subdivision were greater among continents than among local populations, this extension represents an important advance in resolution. Moreover, the fact that linkage disequilibrium among local populations was statistically significant suggests a cautionary note to the genetic epidemiologist considering mixing local populations to fine map disease genes.

Hill and Robertson's (1968) model did not fit the data well. This observation suggests that simple models do not provide a reasonable framework for understanding worldwide linkage disequilibrium at *ALDH2*. Violating the assumptions of the model, *ALDH2* linkage disequilibrium was likely maximal in a finite-sized ancestral population. Further, natural selection may have acted (R.J. Peterson, D. Goldman, and J.C. Long, in prep.), mutation is evident, human population sizes have not been equal (Urbanek et al. 1996), and the hierarchy of human populations is not balanced (Nei and Roychoudhury 1993). This suggests that other simple models, such as Ohta's (1982a,b) partition of the variance of linkage disequilibrium, will also not fit the data well.

The extension of the model of Nei and Li (1973) to two levels provided valuable insights that would have been missed with a one-level partition. However, it can also be concluded that this simple two-level partition was inadequate to fully describe the effects of human demographic history on *ALDH2* linkage disequilibrium. Perhaps the crucial improvement is to model unequal rates of evolution and a realistic human population phylogeny (Nei and Roychoudhury 1993; Urbanek et al. 1996). The emergence of fine-scale haplotype data for many genes in many populations is likely to provide continuing impetus to incorporate population subdivision into coalescence models of linkage disequilibrium (Rannala and Slatkin 1998).

METHODS

Molecular Methods and Population Samples

ALDH2 is defined as the segment of DNA from which *aldehyde dehydrogenase 2* is transcribed (Fig. 1). The six sites analyzed here are a subset of the 12 sites reported by Peterson et al. (1999). The six sites not included in this analysis had only a rare variant, and rare variants are uninformative of the effects of population subdivision on linkage disequilibrium. Site 1 was discovered by M. Stewart (pers. comm.). Sites 2, 3, 5, and 6 were discovered by Peterson et al. (1999). Site 12 is the site that defines the well-known Glu-487–Lys deficiency allele (Yoshida et al. 1984). PCR, restriction enzymes, and SSCP methods were used to genotype the variable sites. Genotypes were collected on a worldwide sample consisting of Africa, 51 Biakans; Asia, 24 Cambodians, 47 Han Chinese, 49 Japanese, 40 South Koreans, 43 Taiwanese, and 50 Black Thai; Europe, 32 Ceph, 41 Finns, 45 Swedes; North America, 51 Cheyenne, 50 Maya, 46 Navajo, 45 Pima; and South America, 49 Karitiana, 44 Rondonian Surui, and 49 Ticuna. Samples were provided and donated by a variety of researchers (Peterson et al. 1999).

Statistical Analyses

For each site, the allele with the higher worldwide frequency was assigned to be the reference allele. Because phase-unknown multi-site genotypes were collected, haplotype states and frequencies were estimated by maximum-likelihood using an expectation-maximization (E-M) method (Dempster et al. 1977). Details of this method, and the associated jackknife standard errors, are presented in Long et al. (1995) and Peterson et al. (1999).

The number of segregating sites in each population ranged from four to six. A contingency table χ^2 test for departure from single-site Hardy-Weinberg expectation (Weir 1996) was applied to each segregating site in each population. Altogether, 68 tests were performed. Four tests had P-values of <5%. These tests lacked independence due to the correlation of alleles among sites. Despite this result, it is reasonable to conclude that this number of departures from Hardy-Weinberg expectation reflects sampling fluctuation under the null hypothesis.

For linkage disequilibrium analysis, the two-site haplotype frequencies were obtained from the six-site haplotype frequencies by summing frequencies of all haplotypes with each specific combination of alleles at the two sites (Long et al. 1995). Linkage disequilibrium in the worldwide data set was partitioned as follows, where D_T, D_W, D_{SC}, and D_{CT} are as defined in the Results. Suppose there are K populations across C continents, and K_c populations on the cth continent. With respect to the total sample, each population has relative size w_i, with

$$\sum_{i=1}^{K} w_i = 1.0$$

and each continent has relative size w_c, with

$$\sum_{c=1}^{C} w_c = 1.0$$

With respect to a continental pooling, each population has relative size

$$\sum_{i=1}^{K_c} w_{ci} = 1.0$$

Following Nei and Li (1973),

$$D_W = \sum_{i=1}^{K} w_i D_i$$

$$D_{SC} = \sum_{c=1}^{C} \sum_{i=1}^{K_c} w_{ci}(p_{ci} - p_c)(q_{ci} - q_c)$$

$$D_{CT} = \sum_{c=1}^{C} w_c(p_c - p_T)(q_c - q_T)$$

Here, p_{ci} is the reference allele frequency at the first site in the ith population on the cth continent, and q_{ci} is the analogous reference allele frequency at the second site. As these equations show, D_{SC} and D_{CT} reflect allele frequency differences among populations (Sinnock and Sing 1972). Because of this diff., D_W, D_{SC}, and D_{CT} were normalized to the total linkage disequilibrium such that $d_W = D_W/D_T$, $d_{SC} = D_{SC}/D_T$ and $d_{CT} = D_{CT}/D_T$. Standard errors were calculated by use of a bootstrap method.

The variance of linkage disequilibrium is D^2 in replicate subpopulations of finite size drawn from a population initially in linkage equilibrium (Hill and Robertson 1968). Because D^2 depends on allele frequency, it is normalized to the variances of allele frequency to obtain the squared correlation coefficient (r^2). In relation to F_{ST}, $E[r^2] = [6(1 - F_{\mathrm{ST}})= 5(1 - F_{\mathrm{ST}})^3 - (1 - F_{\mathrm{ST}}{}^6]$ 15(Hill and Robertson 1968). F-statistics for a two-level partition with equal effects were estimated using the method of Urbanek et al. (1996). The relationship $(1 - F_{\mathrm{ST}}) = (1 - F_{\mathrm{SC}})(1 - F_{\mathrm{CT}})$ (Wright 1978) was used to obtain F_{ST}. Standard errors of the estimates were obtained by use of a jackknife procedure (Weir 1996).

ACKNOWLEDGMENTS

We thank Longina Akhtar for maintaining the Laboratory of Neurogenetics cell lines. Ken Kidd, Su-Jen Tsai, Dr. Chandanayingyong, and Dr. Park kindly provided DNA samples. Mark Stewart provided details on the variant at site 1. Margrit Urbanek, Andrew Bergen, and Jaakko Lappalainen contributed invaluable advice on laboratory techniques and useful discussions. Ken Weiss, Andy Clark, Mark Stoneking, and Henry Harpending provided helpful comments on earlier drafts. We are grateful to two anonymous reviewers for insightful comments. This work was supported by a National Institute of Alcohol Abuse and Alcoholism predoctoral Intramural research training award to R.J.P.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Ajioka, R.S., L.B. Jorde, J.R. Gruen, P. Yu, D. Dimitrova, J. Barrow, E. Radisky, C.Q. Edwards, L.M. Griffen, and J.P. Kushner. 1997. Haplotype analysis of Hemochromatosis: Evaluation of different linkage-disequilibrium approaches and evolution of disease chromosomes. *Am. J. Hum. Genet.* **60:** 1439–1447.

Bowcock, A.M., J.M. Hebert, J.L. Mountain, J.R. Kidd, J. Rogers, K.K. Kidd, and L.L. Cavalli-Sforza. 1991. Study of an additional 58 DNA markers in five human populations from four continents. *Gene Geogr.* **5:** 151–173.

Castiglione, C.M., A.S. Deinard, W.C. Speed, G. Sirugo, H.C. Rosenbaum, Y. Zhang, D.K. Grandy, E.L. Grigorenko, B. Bonne-Tamir, A.J. Pakstis, J.R. Kidd, and K.K. Kidd. 1995. Evolution of haplotypes at the DRD2 locus. *Am. J. Hum. Genet.* **57:** 1445–1456.

Clark, A.G., K.M. Weiss, D.A. Nickerson, S.L. Taylor, A. Buchanan, J. Stengård, V. Salomaa, E. Vartiainen, M. Perola, E. Boerwinkle, and C.F. Sing. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63:** 595–612.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B.* **39:1–38.**

Goldman, D. and M.-A. Enoch. 1990. Genetic epidemiology of ethanol metabolic enzymes: A role for selection. *World Rev. Nutr. Diet* **63:** 143–160.

Harding, R.M., S.M. Fullerton, R.C. Griffiths, J. Bond, M.J. Cox, J.A. Schneider, D.S. Moulin, and J.B. Clegg. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60:** 772–789.

Hästbacka, J., A. de la Chapelle, I. Kaitila, P. Sistonen, A. Weaver, and E. Lander. 1992. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat. Genet.* **2:** 204–211.

Hästbacka, J., A. de la Chapelle, M.M. Mahtani, G. Clines, M.P. Reeve-Daly, M. Daly, B.A. Hamilton, K. Kusumi, B. Trivedi, A. Weaver, A. Coloma, M. Lovett, A. Buckler, I. Kaitila, and E. Lander. 1994. The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell* **78:** 1073–1087.

Hill, W.G. and A. Robertson. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **33:** 54–78.

Hsu, L.C., R.E. Bendel, and A. Yoshida. 1988. Genomic structure of the human mitochondrial aldehyde dehydrogenase gene. *Genomics* **2:** 57–65.

Ikuta, T., S. Szeto, and A. Yoshida. 1986. Three human alcohol dehydrogenase subunits: cDNA structure and molecular and evolutionary divergence. *Proc. Natl. Acad. Sci.* **83:** 634–638.

Jorde, L.B., W.S. Watkins, D. Viskochil, P. O’Connell, and K. Ward. 1993. Linkage disequilibrium in the Neurofibromatosis 1 (NF1) region: Implications for gene mapping. *Am. J. Hum. Genet.* **53:** 1038–1050.

Jorde, L.B., W.S. Watkins, M. Carlson, J. Groden, H. Albertsen, A. Thliveris, and M. Leppert. 1994. Linkage disequilibrium predicts physical distance in the Adenomatous Polyposis Coli region. *Am. J. Hum. Genet.* **54:** 884–898.

Jorde, L.B., M.J. Bamshad, W.S. Watkins, R. Zenger, A.E. Fraley, P.A. Krakowiak, K.D. Carpenter, H. Soodyall, T. Jenkins, and A.R. Rogers. 1995. Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. *Am. J. Hum. Genet.* **57:** 523–538.

Kerem, B.-S., J.M. Rommens, J.A. Buchanan, D. Markiewicz, T.K. Cox, A. Chakravarti, M. Buchwald, and L.-C. Tsui. 1989. Identification of the Cystic Fibrosis gene: Genetic analysis. *Science* **245:** 1073–1080.

Laan, M. and S. Pääbo. 1997. Demographic history and linkage disequilibrium in human populations. *Nat. Genet.* **17:** 435–438.

———. 1998. Mapping genes by drift-generated linkage disequilibrium. *Am. J. Hum. Genet.* **63:** 654–656.

Lewontin, R.C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49:** 49–67.

———. 1995. The detection of linkage disequilibrium in molecular sequence data. *Genetics* **140:** 377–388.

Long, J.C., A. Chakravarti, C.D. Boehm, S. Antonarakis, and H.H. Kazazian. 1990. Phylogeny of human b-globin haplotypes and its implications for recent human evolution. *Am. J. Phys. Anthropol.* **81:** 113–130.

Long, J.C., R.C. Williams, and M. Urbanek. 1995. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56:** 799–810.

Lonjou, C., A. Collins, and N.E. Morton. 1999 Allelic association between marker loci. *Proc. Natl. Acad. Sci.* **96:** 1621–1626.

Nei, M. and W.-H. Li. 1973. Linkage disequilibrium in subdivided populations. *Genetics* **75:** 213–219.

Nei, M. and A.K. Roychoudhury. 1993. Evolutionary relationships of human populations on a global scale. *Mol. Biol. Evol.* **10:** 927–943.

Ohta, T. 1982a. Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proc. Natl. Acad. Sci.* **79:** 1940–1944.

———. 1982b. Linkage disequilibrium with the island model. *Genetics* **101:** 139–155.

Peterson, R.J., D. Goldman, and J.C. Long. 1999. Nucleotide sequence diversity in non-coding regions of ALDH2 as revealed by restriction enzyme and SSCP analysis. *Hum. Genet.* **104:** 177–187.

Raghunathan, L., L.C. Hsu, I. Klisak, R.S. Sparkes, A. Yoshida, and T. Mohandas. 1988. Regional localization of the human genes for aldehyde dehydrogenase-1 and aldehyde dehydrogenase-2. *Genomics* **2:** 267–269.

Rannala, B. and M. Slatkin. 1998. Likelihood analysis of disequilibrium mapping, and related problems. *Am. J. Hum. Genet.* **62:** 459–473.

Sinnock, P. and C.F. Sing. 1972. Analysis of multilocus genetic systems in Tecumseh, MI. II. Considerations of the correlation between nonalleles in gametes. *Am. J. Hum. Genet.* **24:** 393–415.

Smouse, P.E. and J.V. Neel. 1977. Multivariate analysis of gametic disequilibrium in the Yanomama. *Genetics* **85:** 733–752.

Smouse, P.E., J.V. Neel, and W. Liu. 1983. Multiple-locus departures from panmictic equilibrium within and between village gene pools of Amerindian tribes at different stages of agglomeration. *Genetics* **104:** 133–153.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105:** 437–460.

Terwilliger, J.D., S. Zollner, M. Laan, and S. Pääbo. 1998. Mapping genes through the use of linkage disequilibrium generated by genetic drift: "Drift mapping" in small populations with no demographic expansion. *Hum. Hered.* **48:** 138–154.

Thompson, E.A., S. Deeb, D. Walker, and A.G. Motulsky. 1988. The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII Apolipoprotein genes. *Am. J. Hum. Genet.* **42:** 113–124.

Tishkoff, S.A., E. Dietzsch, W. Speed, A.J. Pakstis, J.R. Kidd, K. Cheung, B. Bonné-Tamir, A.S. Santachiara-Benerecetti, P. Moral, M. Krings, S. Pääbo, E. Watson, N. Risch, T. Jenkins, and K.K. Kidd. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271:** 1380–1387.

Tishkoff, S.A., A. Goldman, F. Calafell, W.C. Speed, A.S. Deinard, B. Bonne-Tamir, J.R. Kidd, A.J. Pakstis, T. Jenkins, and K.K. Kidd. 1998. A global haplotype analysis of the Myotonic Dystrophy locus: Implications for the evolution of modern humans and for the origin of Myotonic Dystrophy mutations. *Am. J. Hum. Genet.* **62:** 1389–1402.

Urbanek, M., D. Goldman, and J.C. Long. 1996. The apportionment of dinucleotide repeat diversity in Native Americans and Europeans: A new approach to measuring gene identity reveals asymmetric patterns of divergence. *Mol. Biol. Evol.* **13:** 943–953.

Weber, J.L. and C. Wong. 1993. Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2:** 1123–1128.

Weir, B.S. 1996. *Genetic data analysis II: Methods for discrete population genetic data*. Sinauer, Sunderland, MA.

Wright, S. 1978. Evolution and the genetics of populations. In *Variability within and among natural populations*. Vol. 4. The University of Chicago Press, Chicago, IL.

Yoshida, A., I.-Y. Huang, and M. Ikawa. 1984. Molecular abnormality of an inactive aldehyde–dehydrogenase variant commonly found in Orientals. *Proc. Natl. Acad. Sci.* **81:** 258–261.

Zapata, C., G. Alvarez, and C. Carollo. 1997. Approximate variance of the standardized measure of gametic disequilibrium D′. *Am. J. Hum. Genet.* **61:** 771–774.

Received Janurary 7, 1999; accepted in revised form June 29, 1999.